CONTINUING TECHNICAL AND MECHANICAL CONCERNS RELATING TO IMPLEMENTATION OF REGRESSION ANALYSIS CAP-BASED MODELS

The Commission's recent decision[1] to modify impacts of the regression formulas previously adopted by the Wireline Competition Bureau will temporarily ameliorate, but not resolve, concerns about numerous technical and mechanical flaws in the underlying formulas.

The following comments on these issues are offered as suggestions for resolving some of the more significant technical problems with the formulas; prompt movement to address these concerns is essential given the Commission's indications that it will neither stay application of the caps nor use them only as a trigger pending the completion of much-needed additional corrections. These suggestions are in addition to, and do not waive, any of the significant policy or legal concerns arising with respect to the use of such caps in the first instance.[2]

1. Study Area Boundaries Need Correction

The record shows that the current models suffer from study area boundary errors. The FCC needs to complete a process to produce correct boundary data for all RLECs, while being mindful that producing such data may be a substantial burden for small carriers. As Exhibit 1 shows, the correction of geographic boundary data is likely to produce significant swings in coefficients. This confirms that the model may be subject to wide swings based on correction of only a single study area boundary, and that predictability or certainty will be unobtainable until, at a minimum, this process is complete and coefficients are reset.

2. Census Blocks Must be Correctly Matched to RLEC Study Areas

Once corrected data are in hand, mismatches in mapping of census blocks to study area boundaries must be addressed. Steps must be taken to identify materially incorrect cases, using land area comparisons between corrected study area maps and census block collections. Census blocks that overlap study area boundaries, and census blocks assigned to a study area that does not overlap its area, can be identified. A determination of the correct assignment of these census blocks should be done. This can be achieved without major manual review of maps and data using the Python language with ArcGIS. Once this process produces a list of potential material mismatches, corrections must be made using apportionment, or a more targeted allocation in cases that demonstrate more material concerns.

---

[1] *Connect America Fund*, WC Docket No. 10-90, *High-Cost Universal Service Support*, WC Docket No. 05-337, Sixth Order on Reconsideration and Memorandum Opinion and Order, FCC 13-16 (rel. Feb. 27, 2013) (*Sixth Order on Reconsideration*).

[2] Several of these issues are pending before the 10th Circuit Court of Appeals. *See* Petitioners' Uncited Joint Universal Service Fund Principal Brief, *In re: FCC*, No. 11-9900 (10th Cir., Oct. 23, 2012). The Rural Associations also reserve the right to seek reconsideration or review of actions taken by the Commission in the *Sixth Order on Reconsideration.*

3. The Dependent Variable Should be Cost per Loop

The coefficients display counterintuitive signs in the current models, which should be corrected. This problem has been caused in part by basing the model on total study area cost, rather than on study area average Cost per Loop. The Bureau's use of the total study area cost variable also causes its R-squared statistic to be misleading, showing primarily how well the model fits data of the largest study areas, rather than how it fits data overall. A model to determine conditions of high cost of service per customer needs to assess the R-squared statistic based on Cost per Loop.

4. Independent Variables Should be Restructured

Better estimates might be obtained if independent variables were expressed as ratios rather than as total study area measures. Use of total study area measures tends to bias the model toward the data of larger study areas. Some variables, such as Percent Undepreciated Plant and Density, are already expressed as ratios (although these too require more careful review). Other variants of ratios, such as loops per exchange, road crossings per household, and/or road miles per loop should be reviewed and tested. Variables should be carefully selected based on their effectiveness in the models and on cost causation, not merely based on contribution to the R-squared statistic. With ratios as independent variables, logarithmic transformations may not be needed, although testing is still required in every instance to ensure accuracy and statistical integrity/significance.

5. Independent Variables Should be Chosen with Deliberation

The models should contain deliberately chosen sets of independent variables. Exhibit 2 shows that current variables are not reliable as estimators of quantiles. A reliable estimator of a quantile would have a coefficient significantly different than obtained if the same variable were used to estimate a different quantile. The exhibit shows the coefficients of each variable in the Bureau's models over the full range of possible quantile estimates, from the 10^{th} quantile to the 90^{th} quantile. For example, the second graph in this exhibit shows that the variable *log of loops* would have a value of 0.85 in a model of the 10^{th} quantile, and a value of 0.80 in a model of the 90^{th} quantile. This graph also shows a confidence interval calculated about each coefficient value, which is so wide that the interval at every quantile includes the coefficient value at every other quantile level. Setting aside the question of whether the coefficient of the 90^{th} quantile should be lower than the coefficient of the 10^{th} quantile, this observation suggests that the variable does not help distinguish one quantile from another. Moreover, these graphs show that other variables are also unhelpful in distinguishing a 90^{th} quantile from any other quantile.

Exhibit 3 further demonstrates that failure to engage in more careful choice of variables results in improperly specified quantiles. For six study areas, the Bureau's model of the 90^{th} quantile actually produces ***lower*** CAPEX limits than if a model of the same type

were used to calculate estimates of the 80th quantile, and in some cases lower than estimates of the 60th quantile.

To help remedy such concerns, sequential comparisons should be made between each variable and the dependent variable, and between each variable and the other independent variables. Comparisons would reveal whether each variable relates linearly to the dependent variable or, if non-linearly, what transformation or stratification is needed to justify including the independent variable.

This analysis should identify independent variables which, by their correlation with each other, measure mostly the same causal effects on the dependent variable, leading to some of the "incorrect" signs of coefficients in the current models. For example, loops per exchange and households per square mile are both measures of density, and may not simultaneously contribute to a reliable model. If so, the "weaker" member of such a pair of variables should be dropped.

Variable relationship structures should also be examined. For example, road crossings may increase as roadway miles increase, except in areas where households per roadway mile are very small, suggesting the road crossing variable might be effective only if broken into components.

6. Opportunity for Individual Carriers to Seek Review of Independent Variable Calculations by Reference to Corrected Study Area Boundaries

After study areas are correctly identified and more accurately matched to census blocks, data from various sources are correctly associated with census blocks, and errors in independent variables are analyzed, individual carriers should be entitled, without the need to petition for a waiver, to seek review of the independent variable calculations to identify and resolve outstanding disparities against actual service area data.

7. Geographical Indicators Should be Intuitively Correct

Oftentimes, use of a total cost dependent variable in a model contributes to counterintuitive signs of coefficients of independent variables, because coefficients are influenced more by size than by unit cost. With this in mind, use of Cost per Loop as the dependent variable may help correct counterintuitive geographic indicators, such as the *Alaska v*ariable in the model. If this choice and the more deliberate choice of other independent variables do not clear up counterintuitive signs, indicator variables should be dropped from the model. This could be taken one step at a time, dropping the indicator variable with the weakest correlation first, retesting the model, and then dropping the next weakest variable to achieve intuitively correct indicators.

8. A Study Should Determine Timing of Updates of Independent Variables

The cycles of updates of each of the databases from which independent variables are derived should be studied, and a policy published for how and when periodic updates would be reflected in the cap models. Because standard updates to data could result in significant and entirely unpredictable shifts in support flows, it is essential: (a) to seek proper notice and comment on the use of any such updated information and the potential effects on the volatility of the models; and (b) to phase-in the effects of any such updates in lieu of "flash-cut" changes in support. For example, Exhibit 4 shows the significant and volatile changes in model coefficients that would have occurred between 2007 and 2012 based *only* upon updates to HCLS data, putting aside additional changes that would have occurred as a result of census data changes and updates to other databases from which independent variables were derived.

As another example, while much of the census data used in the models was obtained from the 2010 census, urban area boundaries were obtained from the 2000 census, which designated urban areas as collections of census block groups determined to be urban in character (apparently because of availability). In contrast, the 2010 census constructs urban boundaries by combining census tracts (collections of census block groups) determined in aggregate to be urban. Reliance on census data can be expected to produce potentially material swings in most of the independent variables, even if all other inputs are held constant.

Other independent variables are also subject to their own providers' update cycles. Changes such as these would affect support flows as the caps are periodically reset, even if no other data change. Databases found to be subject to evolving measurement methods should be discarded or remedied in a statistically valid manner.

9. Predictability of Support Resulting From Cap Updates Must be Assured

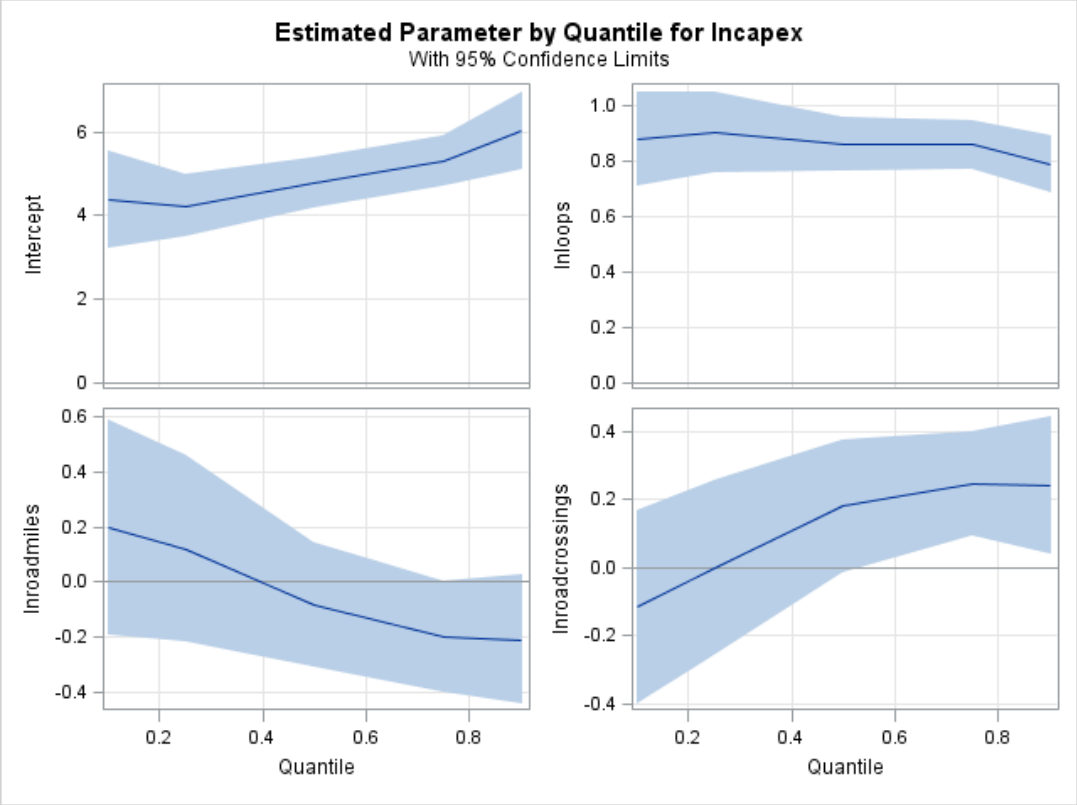
The effects of model updates on cap levels and support payments must be tested based on the cycle of updates to variables. For example, if the model is to be updated in a given year, the Bureau could use HCLS data from prior annual periods to produce a model of the same structure as the benchmark models resulting from these analyses. Any of the independent variables for which the corresponding prior views can be obtained should be included. For example, 2000 census data should be used to test the effects of decennial census changes. This review, among others, is necessary to determine the stability of the caps.

Exhibit 4 underscores the importance of testing for and establishing a more predictable model. As explained earlier, Exhibit 4 shows the significant volatility resulting just from changes to HCLS data. These problematic effects are direct results of very weak independent variables (demonstrated by the non-significant t-statistics associated with the Bureau's models). These problems might only be compounded by updates to other databases and subsequent recalculation of the coefficients and the renewed runs of the models. Once variable choices are made, predictability needs to be confirmed by testing year to year effects of changes to those variables on models.

Exhibit 1
Effects of Data Correction on CAPEX Model Coefficients

Variable	FCC Order Table 3	Revised for Data Correction	% Change
Loops	0.76082	0.78783	-3.4%
Road Miles	-0.14821	-0.20798	-28.7%
Road Crossings	0.21196	0.24044	-11.8%
Count of States	-0.06813	-0.07015	-2.9%
Per Cent Undepreciated Plant	0.03048	0.03069	-0.7%
Density	-0.12701	-0.15783	-19.5%
Exchange Count	0.11668	0.11775	-0.9%
Per Cent Bedrock	-0.08785	-0.07241	21.3%
Soils Difficulty	0.11457	0.11838	-3.2%
Climate	0.09502	0.08864	7.2%
Per Cent Tribal Land	0.00029	0.00048	-39.6%
Per Cent Park Land	0.01702	0.01759	-3.2%
Per Cent Urban	0.00046	0.00058	-20.7%
Alaska	-0.48971	-0.62233	-21.3%
Midwest	0.09783	0.09175	6.6%
Northeast	-0.30917	-0.30902	0.0%
Intercept	6.00019	6.03898	-0.6%

Exhibit 2


Estimated Parameter by Quantile for lncapex
With 95% Confidence Limits
Intercept
lnloops
lnroadmiles
lnroadcrossings
Quantile
Quantile

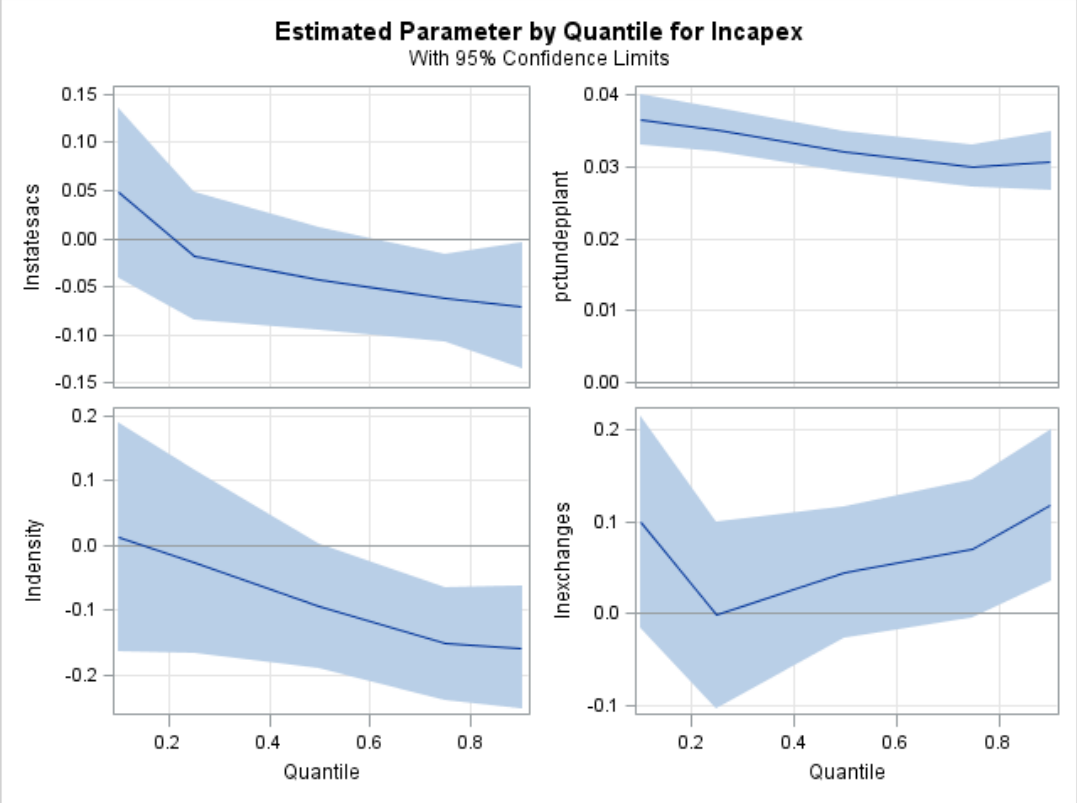
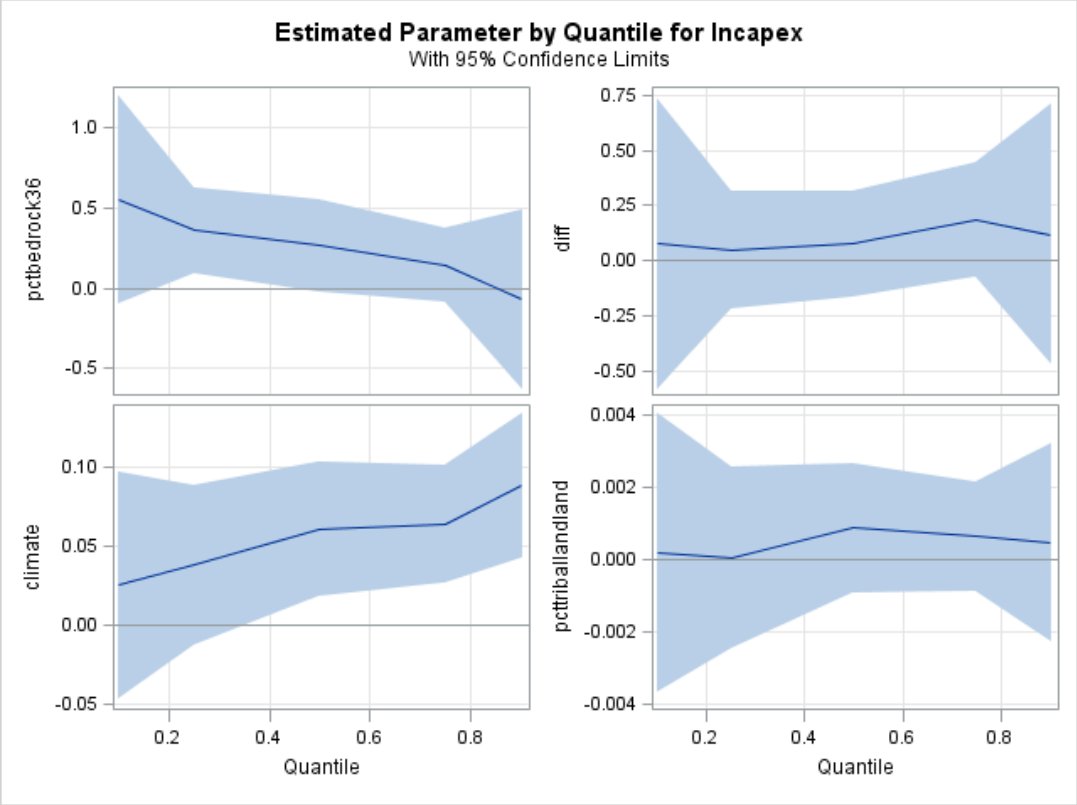

Estimated Parameter by Quantile for lncapex
With 95% Confidence Limits
lnstatesacs
pctundepplant
lndensity
lnexchanges
Quantile
Quantile

Exhibit 2 (Continued)



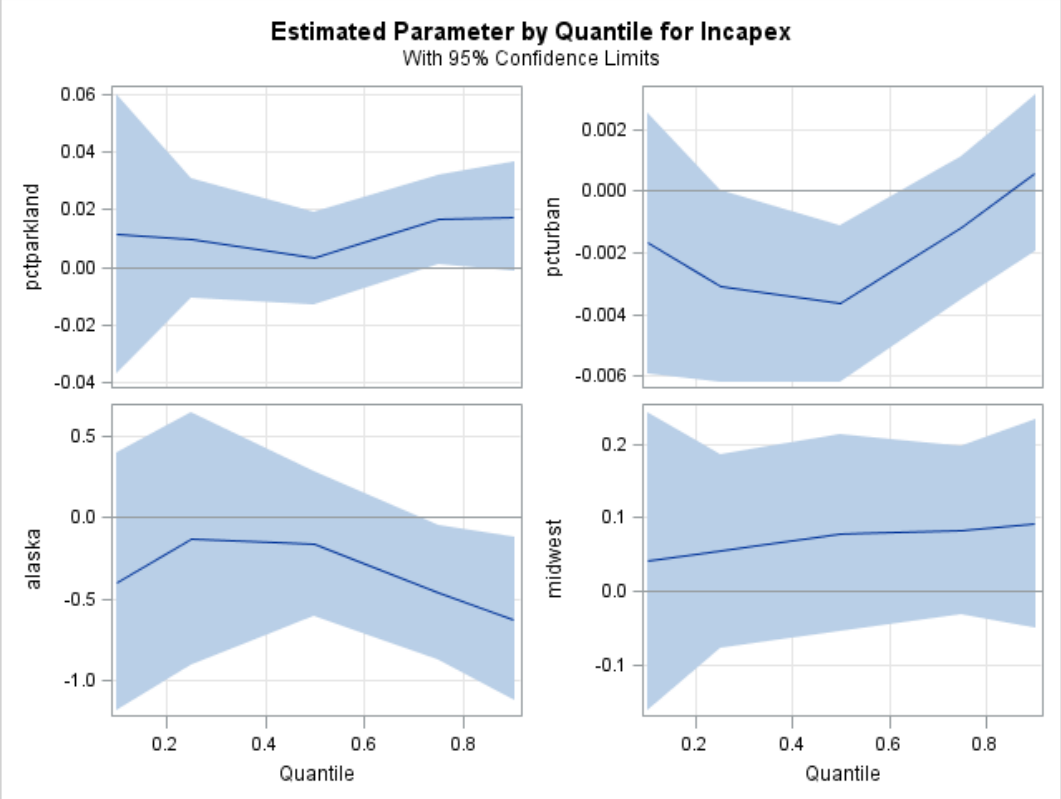


Exhibit 2 (Continued)

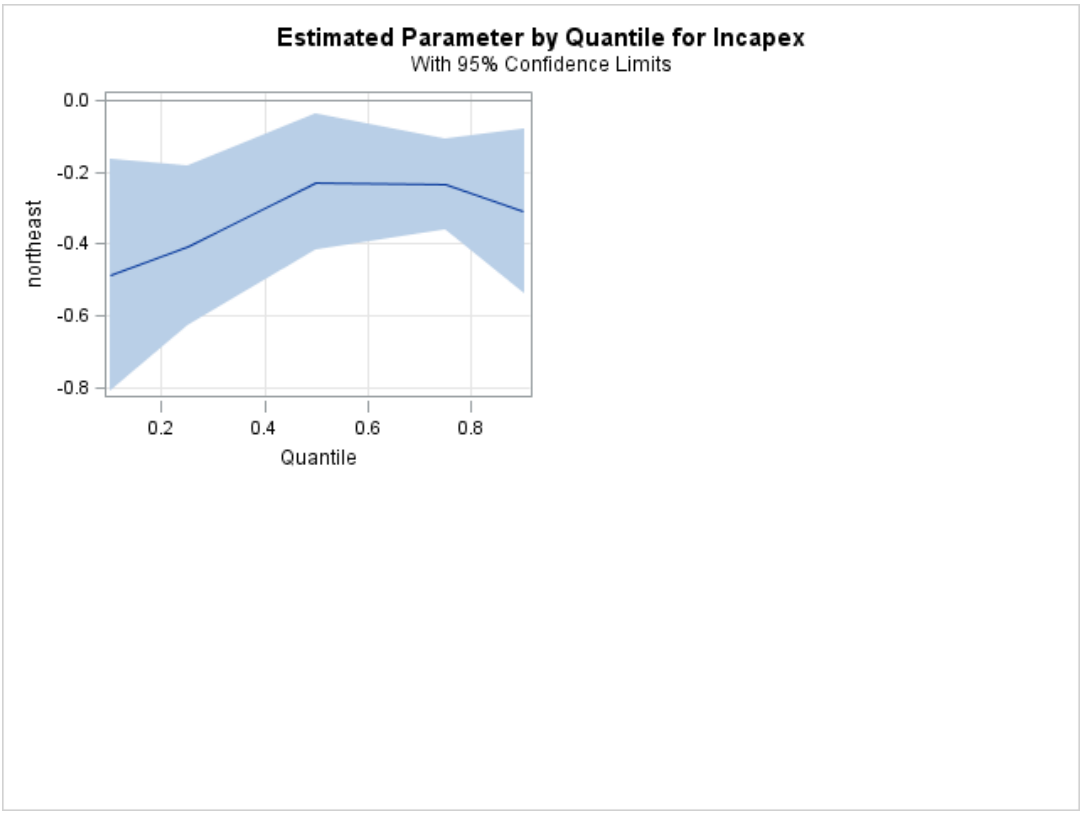
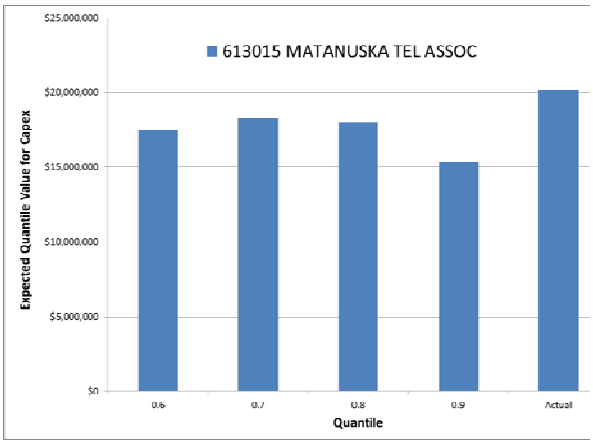
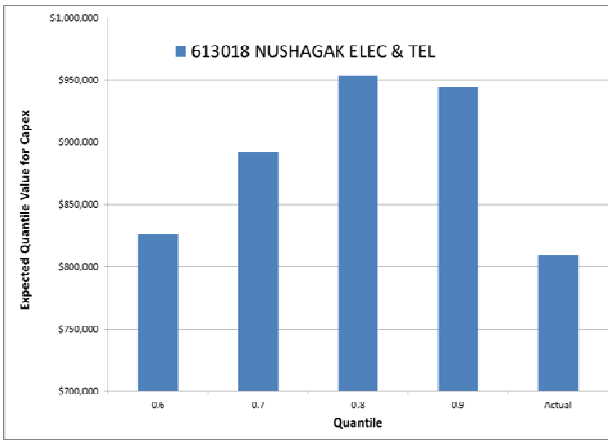
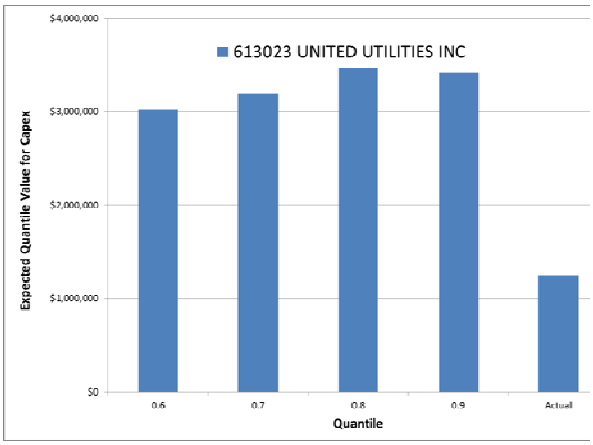
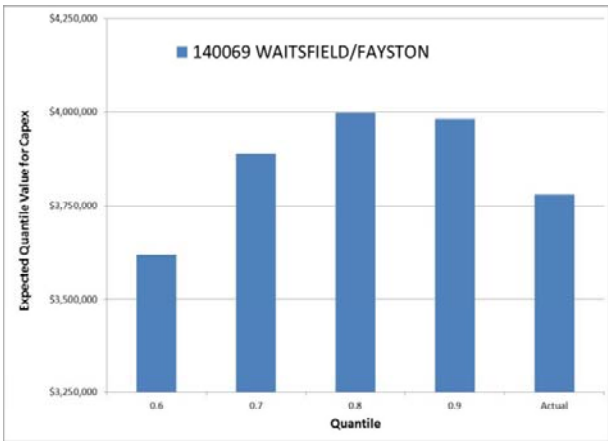


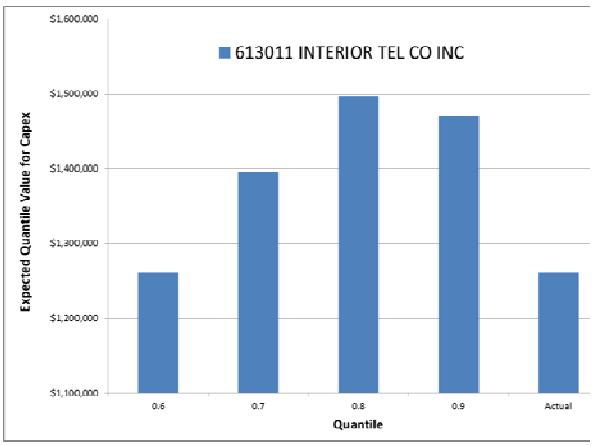
Exhibit 3
Non-Monotonic Quantiles (90th Quantile < Lower Quantiles)











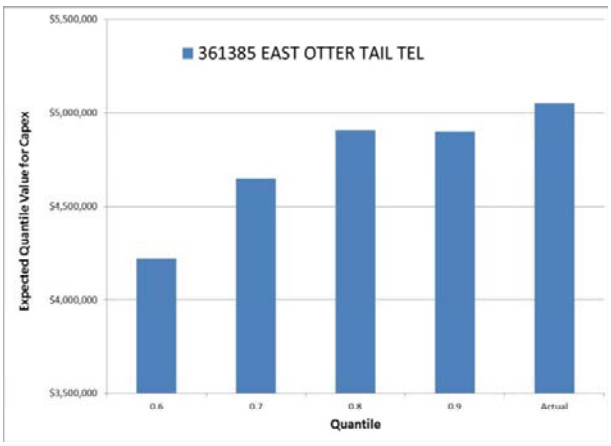


Exhibit 4
Effects of Annual Updates of HCLS Data on Models

CAPEX Model					
Independent Variable	HCLS Support Year 2011	2010	2009	2008	2007
lnloops	0.7878	0.7508	0.6713	0.6509	0.6488
lnroadmiles	-0.2080	-0.1556	0.0199	-0.0018	-0.0056
lnroadcrossin	0.2404	0.2043	0.1216	0.1838	0.2018
lnstatesacs	-0.0702	-0.0521	-0.0240	-0.0486	-0.0408
pctundepplant	0.0307	0.0293	0.0315	0.0284	0.0272
lndensity	-0.1578	-0.1502	-0.0875	-0.0761	-0.0734
lnexchanges	0.1178	0.1310	0.1226	0.1376	0.1247
pctbedrock36	-0.0724	-0.1630	-0.1943	-0.0256	0.0075
Diff	0.1184	0.0828	0.1437	0.1182	0.0921
climate	0.0886	0.0905	0.1075	0.1107	0.0994
pcttriballand	0.0005	0.0009	0.0019	0.0016	0.0020
pctparkland	0.0176	0.0186	0.0151	0.0166	0.0161
pcturban	0.0006	0.0002	0.0019	0.0017	0.0023
alaska	-0.6223	-0.6695	-0.3610	-0.2369	-0.2730
midwest	0.0918	0.0674	0.0841	0.1205	0.0781
northeast	-0.3090	-0.2814	-0.1199	-0.0179	-0.0215

OPEX Model					
Independent Variable	HCLS Support Year 2011	2010	2009	2008	2007
lnloops	0.5958	0.6972	0.7040	0.7008	0.7075
lnroadmiles	-0.2470	-0.3329	-0.3369	-0.2968	-0.2659
lnroadcrossin	0.2723	0.2909	0.3104	0.2450	0.2148
lnstatesacs	-0.0778	-0.0677	-0.0703	-0.0709	-0.0755
pctundepplant	0.0077	0.0058	0.0080	0.0097	0.0096
lndensity	-0.1276	-0.1896	-0.1614	-0.1455	-0.1381
lnexchanges	0.1250	0.1239	0.0943	0.1401	0.1134
pctbedrock36	0.2789	0.1957	0.3985	0.2699	0.2354
diff	0.1141	0.0910	0.0745	0.1688	0.2482
climate	0.1351	0.1314	0.1233	0.1384	0.1378
pcttriballand	0.0019	0.0016	0.0027	0.0021	0.0030
pctparkland	0.0064	0.0042	0.0073	0.0016	0.0004
pcturban	0.0025	0.0028	-0.0007	0.0006	-0.0005
alaska	0.2989	0.2301	0.0762	0.4945	0.3623
midwest	0.1338	0.1404	0.1408	0.1910	0.1628
northeast	0.0149	0.0068	0.0167	0.0329	0.0347